WHEN DOES IT PAY TO BREAK THE MATCHES FOR ANALYSIS OF A MATCHED PAIRS DESIGN?

Henry S. Lynn, Charles E. McCulloch

Biometrics Unit, Cornell University, Ithaca, NY 14853

BU-1042-MB[1] January, 1990

SUMMARY

Two methods of analyses are compared to estimate the treatment effect of a comparative study where each treated individual is matched with a single control at the design stage. The usual matched pairs analysis accounts for the pairing directly in its model, whereas regression adjustment ignores the matching but instead models the pairing using a set of covariates. For a normal linear model, the estimated treatment effect from the matched pairs analysis (paired t-test) is more efficient. For a Bernoulli logistic model, matched pairs analysis performs better when the sample size was small, but is inferior to logistic regression for large sample sizes.

---

[1]Technical Report BU-1042-MA in the Biometrics Unit Series.

1. INTRODUCTION

In many comparative studies, researchers may decide at the design stage to form pairs by matching exactly on some not so readily quantifiable variable; e.g., sibship or neighborhood of residence, but in the subsequent analysis it is not always clear what statistical analyses they should select. The usual approach is to employ a matched pairs analysis. This method will yield an unbiased or asymptotically unbiased estimate of the "treatment" effect, although the estimate may have a larger variance than, say, that obtained from a regression analysis which attempts to model the pairing using some set of covariates. The latter approach sacrifices the unbiasedness of the estimate in an attempt to gain higher precision. The purpose of this paper is then to determine, for both the normal linear model and Bernoulli logistic model, which of the above two analyses is preferable by quantitatively assessing the trade off between the loss in accuracy and increase in efficiency involved in modelling the pairing.

Prentice (1976), and Breslow and Day (1980), using examples of case-control studies, have shown that a logistic regression analysis which failed to account for "important" covariates will tend to bias the regression estimates, whereas an analysis that included "redundant" covariates will inflate the variances of the estimates. In particular, Prentice (1976) cites a matched pairs study relating post-menopausal estrogen exposure on endometrical cancer, and recommends an analysis without retaining the pairing since pairs members were not intrinsically similar in respects other than those indicated by the matching variables. However, their

findings were more qualitative, and do not indicate how "important" the covariates need to be; i.e., how adequate the covariates are in explaining the pairing before regression analysis becomes more effective than a matched pairs analysis. Extensive work on the problem of omitted covariates in general linear models has been done by Gail et al. (1984, 1988), in which the covariates are treated as random variables. Many authors have also examined the pros and cons of matching versus regression adjustment, although their comparisons are restricted to the case when the pairing has been fully modelled by the covariates. Others have approached the problem from an experimental viewpoint, comparing the efficiencies of unmatched versus matched designs. For an overview of some of the results, see Rubin (1973), McKinlay (1977), Kupper et al. (1981), and Greenland (1986).

In the following comparisons between matching and regression adjustment, it will be assumed that we have a comparative study involving a control group and a "treatment" group. There will be a total of m pairs and n = 2m subjects, where each subject from the treatment group is paired with a single subject from the control group.

2. NORMAL LINEAR MODEL

Suppose that the true model is given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\gamma + \mathbf{e} \quad , \tag{1}$$

where $\mathbf{Y}$ is arranged such that (y_i, y_{i+m}), $1 \le i \le m$, are the responses of the ith pair, $\boldsymbol{\beta}' = (\beta_1, \beta_2, ..., \beta_s)$ and $\gamma' = (\gamma_1, ..., \gamma_t)$ represent vectors of

parameters, $\mathbf{X} = \begin{bmatrix} 1 & 1 \mid \mathbf{X}_1 \\ 1 & -1 \mid \mathbf{X}_2 \end{bmatrix}_{n \times s}$ and $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix}_{n \times t}$ are matrices of arbitrary fixed constants, and $E(\mathbf{e}) = \mathbf{0}$ and $Var(\mathbf{e}) = \sigma_e^2 \mathbf{I}_n$. This model therefore assumes that the "intra-match" correlation can be explained by a finite set of covariates within a linear regression model format.

In the following, $\mathbf{X}$ will represent the covariates used to model the pairing and $\mathbf{Z}$ will represent the covariates omitted from the model. Moreover, $\mathbf{X}_1$ and $\mathbf{Z}_1$ are the covariates for the treatment group whereas $\mathbf{X}_2$ and $\mathbf{Z}_2$ are the covariates for the control group, and $\mathbf{X}_1 = \mathbf{X}_2$ and $\mathbf{Z}_1 = \mathbf{Z}_2$ because of exact matching. For the model stated in (1), the object of inference will be the treatment effect, defined as $2\beta_2$ for simplicity of analysis.

2.1) Matched Pairs Analysis.

The matched pairs analysis is the usual analysis for two paired samples, with the true model rewritten as

$$\mathbf{Y} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \boldsymbol{\varepsilon} \quad , \text{ where } \boldsymbol{\varepsilon} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \beta_3 \\ \vdots \\ \beta_s \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix} \gamma + \mathbf{e} \ . \tag{2}$$

If we define the pairwise differences,

$$y_i - y_{m+i} = 2\beta_2 + (\varepsilon_i - \varepsilon_{m+i}) \quad \text{for } i = 1, 2, \ldots, m \ ,$$

the treatment effect is then estimated by $2\hat{\beta}_2 \equiv \frac{1}{m} \sum_{i=1}^{m} (y_i - y_{m+i})$, which is

unbiased since $E(\widehat{2\beta_2}) = 2\beta_2 + \frac{1}{m}\sum_{i=1}^{m} E(\varepsilon_i - \varepsilon_{m+i}) = 2\beta_2$. Also

$$Var(\widehat{2\beta_2}) = \frac{1}{m} Var(\varepsilon_i - \varepsilon_{m+i}) = \frac{1}{m} Var(e_i - e_{m+i}) = \frac{2}{m}\sigma_e^2 , \qquad (3)$$

where σ_e^2 is estimated by $\widehat{\sigma}_e^2$, which equals half the value of the estimated variance of the pairwise differences.

2.2) Regression Adjustment.

Suppose that in our study we have measured **X**, a subset of the covariates that determine the pairing. We now decide to break the matches in the analysis and instead model the pairing by regressing **Y** on **X**. We adopt a working model under which

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} , \qquad (4)$$

where $E(\boldsymbol{\xi}) = \mathbf{0}$ and $Var(\boldsymbol{\xi}) = \sigma_\xi^2 \mathbf{I}_n$. Note that under the true model $\boldsymbol{\xi} = \mathbf{Z}\gamma + \mathbf{e}$, such that $E(\boldsymbol{\xi}) = \mathbf{Z}\gamma$ and $Var(\boldsymbol{\xi}) = \sigma_e^2\mathbf{I}_n$. The usual least squares analysis then gives $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ with $Var(\widehat{\boldsymbol{\beta}}) = \sigma_\xi^2(\mathbf{X}'\mathbf{X})^{-1}$. Observe that the above is just an underfitted model of the true model in (1), with **Z** being the omitted covariates. In general this would mean that $\widehat{\boldsymbol{\beta}}$ is biased since

$$E(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\gamma) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\gamma .$$

However in this case $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\begin{bmatrix} \mathbf{1}' & \mathbf{1}' \\ \mathbf{1}' & -\mathbf{1}' \\ \mathbf{X}_1' & \mathbf{X}_2' \end{bmatrix}\begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix}\gamma$

$$= \beta + \begin{bmatrix} 2m & 0 & 1'(X_1 + X_2) \\ 0 & 2m & 0' \\ (X_1 + X_2)'1 & 0 & X_1'X_1 + X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} 1' & 1' \\ 1' & -1' \\ X_1' & X_2' \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \gamma$$

$$= \beta + \begin{bmatrix} C_{11} & 0 & C_{12} \\ 0 & \frac{1}{2m} & 0' \\ C_{21} & 0 & C_{22} \end{bmatrix} \begin{bmatrix} 1' & 1' \\ 1' & -1' \\ X_1' & X_2' \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \gamma,$$

where $C = \begin{bmatrix} 2m & 1'(X_1 + X_2) \\ (X_1 + X_2)'1 & X_1'X_1 + X_2'X_2 \end{bmatrix}^{-1}$.

This implies that $E(\hat{\beta}_2) = \beta_2 + \frac{1}{2m}(1'Z_1 - 1'Z_2)\gamma = \beta_2$, and thus $2\hat{\beta}_2$ is again an unbiased estimator of the treatment effect because $Z_1 = Z_2$ due to exact matching.

Therefore to compare the effectiveness of the matched pairs analysis with regression adjustment, we consider the variance of $2\hat{\beta}_2$ since $2\hat{\beta}_2$ is unbiased for both models. Specifically, we will use the length of a 95% confidence interval of $2\hat{\beta}_2$ as our criterion of comparison. Our working model (4) implies using $(X'X)^{-1}$ above that $\mathrm{Var}(\hat{\beta}_2) = \frac{\sigma_\xi^2}{2m}$, and hence

$$\mathrm{Var}(2\hat{\beta}_2) = \frac{2\sigma_\xi^2}{m}, \tag{5}$$

where σ_ξ^2 is estimated by $\hat{\sigma}_\xi^2$, the residual mean square from the regression of Y on X. Thus, if we define λ as the ratio of the lengths of the 95% confidence intervals of $2\hat{\beta}_2$ from the regression adjustment over

the matched pairs analysis, then using (3) and (5) we have

$$\lambda = \frac{2\, t_{2m-s}\sqrt{\frac{2}{m}\hat{\sigma}_\xi^2}}{2\, t_{m-1}\sqrt{\frac{2}{m}\hat{\sigma}_e^2}}, \tag{6}$$

where t_ν is the upper 0.025 percentage point of the t distribution with ν degrees of freedom.

In order to determine the behavior of λ we will now evaluate $\hat{\sigma}_e^2$ and $\hat{\sigma}_\xi^2$ in terms of their unobservable theoretical values. According to the true model in (1), $\hat{\sigma}_e^2$ is the residual mean square from the regression of **Y** on **X** and **Z**. Comparing (1) with (4) we find that $\boldsymbol{\xi} = \mathbf{Z}\gamma + \mathbf{e}$, and thus

$$(n-s)\,\hat{\sigma}_\xi^2 = SSR(\mathbf{Z} \mid \mathbf{X}) + (n-s-t)\,\hat{\sigma}_e^2, \tag{7}$$

where

$$SSR(\mathbf{Z} \mid \mathbf{X}) = \mathbf{Y}'(\mathbf{PZ})((\mathbf{PZ})'\mathbf{PZ})^{-1}(\mathbf{PZ})'\mathbf{Y} \tag{8}$$

is the sequential sum of squares of **Z** given that **X** has already been fitted, and **P** is defined to be equal to $\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Now observe that $\mathbf{PY} = \mathbf{P}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}) = \mathbf{P}\boldsymbol{\xi}$, and since **P** is idempotent and symmetric this implies that

$$\mathbf{Y}'\mathbf{PZ} = (\mathbf{Y}'\mathbf{P})\mathbf{PZ} = (\mathbf{P}\boldsymbol{\xi})'\mathbf{PZ}. \tag{9}$$

Substituting (9) into (8), we obtain

$$SSR(\mathbf{Z} \mid \mathbf{X}) = (\mathbf{P}\boldsymbol{\xi})'(\mathbf{PZ})((\mathbf{PZ})'\mathbf{PZ})^{-1}(\mathbf{PZ})'\mathbf{P}\boldsymbol{\xi}, \tag{10}$$

which means that $SSR(\mathbf{Z} \mid \mathbf{X})$ is also the regression sum of squares of the regression of $\mathbf{P}\boldsymbol{\xi}$ on **PZ**. Applying this result, we can then rewrite (7) as

$$\hat{\sigma}_e^2 = \frac{n-s}{n-s-t}(1-R^2)\hat{\sigma}_\xi^2, \tag{11}$$

where R^2 is the coefficient of determination of the regression of $\mathbf{P}\boldsymbol{\xi}$ on **P Z**.

Consequently, on substituting (11) into (6) we obtain

$$\lambda = \frac{t_{2m-s}}{t_{m-1}\sqrt{1-R_a^2}},$$

where $1 - R_a^2 \equiv \frac{n-s}{n-s-t}(1-R^2)$.

We use the notation R_a^2 since the quantity defined is similar to an adjusted R^2. In fact, if $\boldsymbol{\xi}$ were observable so that we could fit a model of the form $\boldsymbol{\xi} = \mathbf{Z}\gamma + \mathbf{e}$, then it can be shown that

$$\hat{\sigma}_e^2 = (1 - R_a^2)\hat{\sigma}_\xi^2,$$

where R_a^2 *is* the adjusted R^2 for the regression of $\boldsymbol{\xi}$ on $\mathbf{Z}$.

To illustrate the efficacy of the matched pairs analysis over regression adjustment, consider Table 1 which lists the values of λ for the simple case when s = 3 and t = 1; i.e. when there is one known covariate and one omitted covariate. Observe that the gain in precision for the regression analysis, due to its larger degrees of freedom, diminishes with increasing sample sizes and values of R_a^2, and is never significantly more efficient than the matched pairs analysis for any reasonable sample size and value of R_a^2. For example, if m = 30, then λ = 1.03 when R_a^2 = 0.10. Intuitively, this means that the measured covariate needs to explain more than 90% of the variation of the regression of $\mathbf{P}\boldsymbol{\xi}$ on $\mathbf{PZ}$ (or $\boldsymbol{\xi}$ on $\mathbf{Z}$) before regression adjustment can be as effective as the matched pairs analysis. The behavior of λ follows similarly when there are more covariates; i.e., for larger values of s and t. In summary, the above results indicate that the matched pairs analysis is preferable for the normal linear model.

Table 1. Ratio of the length of a 95% confidence interval using a regression adjustment over a matched pairs analysis.

	R_a^2					
m	0.0	0.1	0.3	0.5	0.7	0.9
5	.85	.90	1.02	1.20	1.56	2.70
10	.93	.98	1.11	1.32	1.70	2.95
30	.98	1.03	1.17	1.39	1.79	3.10
50	.99	1.04	1.18	1.40	1.80	3.13
100	.99	1.05	1.19	1.41	1.81	3.14

3. BERNOULLI LOGISTIC MODEL

Suppose now that $y_1, \ldots, y_n$ are independent binary random variables with densities $h_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$, and the true model is given by

$$\text{logit}(\pi_i) = \ln \frac{\pi_i}{1-\pi_i} = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\gamma, \tag{12}$$

so $\pi_i = E(y_i) = \Pr\{y_i = 1\} = \dfrac{1}{1 + e^{-\mathbf{x}_i'\boldsymbol{\beta} - \mathbf{z}_i'\gamma}}$,

$\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_s)$, $\gamma' = (\gamma_1, \ldots, \gamma_t)$,

$\mathbf{x}_i'$ is the ith row of $\mathbf{X} = \begin{bmatrix} 1 & 1 \mid \mathbf{X}_1 \\ 1 & -1 \mid \mathbf{X}_2 \end{bmatrix}_{n \times s}$, and

$\mathbf{z}_i'$ is the ith row of $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 \\ \\ \mathbf{Z}_2 \end{bmatrix}_{n \times t}$, $i = 1, 2, \ldots, n$.

As in Section 2, $\mathbf{X}_1$ and $\mathbf{Z}_1$ are the covariates for the treatment group whereas $\mathbf{X}_2$ and $\mathbf{Z}_2$ are the covariates for the control group. Furthermore, $\mathbf{X}_1 = \mathbf{X}_2$ and $\mathbf{Z}_1 = \mathbf{Z}_2$ because of exact matching, and $2\beta_2$ will again be used to measure the treatment effect.

3.1) Matched Pairs Analysis.

For a matched pairs design, the typical quantity of interest is the odds ratio, Ψ, or the log odds ratio $\tau = \ln\Psi$. According to model (12) we have $\Psi = e^{2\beta_2}$ and $\tau = 2\beta_2$. Hence, if we let $n_{10} \equiv$ number of pairs with a "1" for the treated subject and a "0" for the control, $n_{01} \equiv$ number of pairs with a "0" for the treated subject and a "1" for the control, and $N_d = n_{10} + n_{01} \equiv$ number of discordant pairs, then according to Breslow and Day (1981) the usual estimator for τ is

$$\hat{\tau}_{ML} = \ln\hat{\psi} = \ln\left(\frac{n_{10}}{n_{01}}\right) = \ln\left(\frac{n_{10}}{N_d - n_{10}}\right), \tag{13}$$

which is the maximum likelihood estimator of the log odds ratio.

In order to evaluate the efficiency of $\hat{\tau}_{ML}$ and subsequently compare it with the estimator from the logistic regression analysis, we need to first obtain the expectation and variance of $\hat{\tau}_{ML}$. Applying a Taylor's series approximation to (13), it can be shown that

$$\mathrm{Var}\,(\hat{\tau}_{ML} \mid N_d) \cong \frac{(\psi+1)^2}{N_d\psi}, \text{ and} \tag{14}$$

$$E(\hat{\tau}_{ML} \mid N_d) \cong \tau + \frac{(\psi+1)(\psi-1)}{2N_d\psi}. \tag{15}$$

It follows from (15) that

$$E(\hat{\tau}_{ML}) = E(E(\hat{\tau}_{ML} \mid N_d)) \cong \tau + \frac{(e^{\tau}+1)(e^{\tau}-1)}{2e^{\tau}} E\left(\frac{1}{N_d}\right), \tag{16}$$

and using (14) and the first term of $E(\hat{\tau}_{ML} \mid N_d)$ in (15) we then find that

$$\mathrm{Var}\,(\hat{\tau}_{ML}) = E(\mathrm{Var}\,(\hat{\tau}_{ML} \mid N_d)) + \mathrm{Var}\,(E(\hat{\tau}_{ML} \mid N_d))$$

$$\cong E\left(\mathrm{Var}\left(\hat{\tau}_{ML}|N_d\right)\right) \cong \frac{\left(e^{\tau}+1\right)^2}{e^{\tau}} E\left(\frac{1}{N_d}\right) . \qquad (17)$$

Since the square of the bias is of order $E\left(\frac{1}{N_d}\right)^2$, it is negligible with respect to the variance and asymptotically we have

$$\mathrm{MSE}(\hat{\tau}_{ML}) \cong \frac{\left(e^{\tau}+1\right)^2}{e^{\tau}} E\left(\frac{1}{N_d}\right) . \qquad (18)$$

3.2) Logistic Regression.

Analogous to the situation described in Section 2.2, we assume here that, **Z**, a subset of the covariates has been omitted, thus while the true model is (12), we are using a working model in which $y_1, \ldots, y_n$ are assumed to have densities $f_i(y_i, \boldsymbol{\beta}) = P_i^{y_i}(1-P_i)^{1-y_i}$, where $P_i = \dfrac{1}{1+e^{-x_i'\boldsymbol{\beta}}}$ and $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_s)$.

Applying a result in Royall's (1986, p.222) paper, we then find that $\hat{\boldsymbol{\beta}}$, the maximum likelihood estimator of $\boldsymbol{\beta}$, is a consistent estimator of the root of the likelihood equation, $\boldsymbol{\beta}^{o}$. For large n, $\boldsymbol{\beta}^{o}$ is approximately the solution of $\sum_{i=1}^{n} (\pi_i - P_i)x_i' = 0$, where the π_i's are the true probabilities as defined by the model in (12).

Furthermore, if we define $P_i^{o} = \dfrac{1}{1+e^{-x_i'\boldsymbol{\beta}^{o}}}$,

$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 \end{bmatrix}$, where $\mathbf{V} = \mathrm{diag}\{\pi_i(1-\pi_i)\}$ for i = 1, 2, ..., n, and

$$\mathbf{V}^{o}=\begin{bmatrix} \mathbf{V}_1^{o} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2^{o} \end{bmatrix} \text{ , where } \mathbf{V}^{o}=\text{diag}\{P_i^{o}(1-P_i^{o})\} \text{ for } i = 1, 2, ..., n,$$

then (Royall, p.222, 1986)

$$\sqrt{n}\,(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}\left(\boldsymbol{\beta}^{o}-\boldsymbol{\beta}\,,\,\Sigma\right), \tag{19}$$

where $\Sigma = n\,(\mathbf{X}'\mathbf{V}^{o}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}\mathbf{X})\,(\mathbf{X}'\mathbf{V}^{o}\mathbf{X})^{-1}$.

Therefore, on applying equation (19), we conclude that

$$\text{Bias}(2\hat{\beta}_2) = 2(\beta_2^{o}-\beta_2), \tag{20}$$

and

$$\text{Var}(2\hat{\beta}_2) = 4\Sigma_{22}\,, \tag{21}$$

where Σ_{22} is the second diagonal element of Σ. .

3.3) Results and Discussion.

Recall that $\hat{\tau}_{ML}$ is the ML estimator of the treatment effect from the matched pairs analysis. Now let $\hat{\tau}_{logit}$ be the corresponding estimator obtained using logistic regression, then from (20) and (21) we find that asymptotically,

$$\text{MSE}(\hat{\tau}_{logit}) = 4\{\Sigma_{22}+(\beta_2^{o}-\beta_2)^2\}. \tag{22}$$

To compare the matched pairs analysis with logistic regression, we consider the simple case when there is only one known covariate and one omitted covariate; i.e. the true model is given by

$$\text{logit}(\pi_i) = \beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\gamma z_i \,,\; 1 \le i \le n.$$

The comparison was investigated in terms of the asymptotic performances of $\hat{\tau}_{ML}$ and $\hat{\tau}_{logit}$ via equations (18) and (22). The quantity $E\left(\frac{1}{N_d}\right)$ in (18) was approximated either using $1/E(N_d)$ or through a normal approximation for $Pr(N_d = k)$, where k=1,2,...,m (the case k=0 is omitted). Since the mean square errors are a complicated function of the parameters and the design, we calculated them for a 2^7 arrangement of the following factors:

$m \equiv$ number of pairs,

$x_3 \equiv$ known covariate vector,

$\beta_1 \equiv$ intercept,

$\beta_2 \equiv$ regression coefficient for considering the treatment effect,

$\beta_3 \equiv$ regression coefficient for the known covariate,

$\gamma \equiv$ regression coefficient for the omitted covariate, and

$\rho \equiv$ correlation between x_3 and z (the omitted covariate vector).

ρ was chosen as a measure of effectiveness of pairing since, intuitively, if the correlation is high, the information in Z is redundant and the pairing will be ineffective. This corresponds to the measure R_a^2 of Section 2.2 in the following way. We expect that the correlation between **Y** and **X** to be greater than **Y** and **Z**, (included covariates more effective than the omitted ones). In such a case, for the normal, linear model, R_a^2 is a monotonically decreasing function of ρ.

Each factor was inspected at two levels. In particular, x_3 was chosen to be the standardized versions of the following two vectors:

(1, 1, ..., 1, −1, −1, ..., −1, 1, 1, ..., 1, −1, −1, ..., −1) , and

$(1, 2, ..., \frac{m}{2}, 1, 2, ..., \frac{m}{2}, 1, 2, ..., \frac{m}{2}, 1, 2, ..., \frac{m}{2})$

(Note that z was also standardized in the computer trial.)

Based upon the asymptotic calculations, $MSE(\hat{\tau}_{logit})$ was uniformly smaller than $MSE(\hat{\tau}_{ML})$, which suggests that logistic regression is the preferred analysis for large sample sizes. Also, of the seven factors considered, it was found that β_1 and x_3 had a much smaller influence on the mean square errors of $\hat{\tau}_{ML}$ and $\hat{\tau}_{logit}$.

The next step was to determine the extent to which the asymptotic results apply for finite sample sizes. Retaining only the five factors that were deemed important in the calculations, a 2^3 by 3^2 simulation experiment was carried out and mean square errors were simulated using GAUSS (Edlefsen and Jones, 1986). (See Appendix 1). For each of the 72 runs, 3000 replications were performed, and the different factor levels chosen were:

$m = 30, 60, 100,$ $\beta_2 = 0.4, 1.2,$ $\rho = 0.2, 0.8,$

$\gamma = 0, 0.5, 1.2,$ $\beta_3 = 0.3, 1.2,$

(x_3 was set to be the standardized form of $(1, 1, ..., 1, -1, -1, ..., -1, 1, 1, ..., 1, -1, -1, ..., -1)$, and β_1 was set to zero.)

As $\hat{\tau}_{ML}$ can give rise to serious bias for small sample sizes, an additional matched pairs estimator was included in the comparison with $\hat{\tau}_{logit}$. This was $\hat{\tau}_H = \ln\left(\frac{n_{10}+0.5}{n_{01}+0.5}\right)$, which seemed to perform better in Jewell's (1984) comparison of several sample estimators of the log odds ratio.

Tables 2 and 3 indicate the relative MSE performance of $\hat{\tau}_{ML}$ and $\hat{\tau}_{H}$ to $\hat{\tau}_{logit}$ by listing the values of r_{ML} and r_{H} for each of the 72 runs, where

r_{ML} is the median of 3000 replicated values of $\dfrac{(\hat{\tau}_{logit} - 2\beta_2)^2}{(\hat{\tau}_{ML} - 2\beta_2)^2}$, and

r_{H} is the median of 3000 replicated values of $\dfrac{(\hat{\tau}_{logit} - 2\beta_2)^2}{(\hat{\tau}_{H} - 2\beta_2)^2}$.

Medians were used as a robust measure since forming ratios sometimes produced extreme values.

1) Most of the values of r_{ML} and r_{H} are less than 1, which indicate that logistic regression performs better, with the exception when ρ is small and γ is large. This agrees with intuition since for our regression model, the smaller ρ is the less "information" about the omitted covariate is being captured, and the larger γ is the more bias we expect in estimating the treatment effect.

2) The most reliable factor determining the behavior of r_{ML} and r_{H} turns out to be the sample size. As the number of pairs increases, the values of r_{ML} and r_{H} will eventually drop below 1. This holds even for the anomaly observed for the small ρ large γ where further simulation showed that more than 200 pairs were needed.

3) In general as gamma increases, r_{ML} and r_{H} also increase.

4) As β_2 increases both r_{ML} and r_{H} decrease, while as β_3 increases both r_{ML} and r_{H} increase, although these two trends seemed reversed when ρ is small and γ is large.

5) r_{ML} and r_{H} do not seem to decease as ρ increases unless γ is large.

We also note that overall there are no sharp differences between the behavior of r_{ML} and r_{H}, aside from the observation that, for small sample sizes, increasing β_3 causes r_H to decrease when β_2 is large, and increasing γ causes r_H to decrease when β_2 and ρ are large.

4. CONCLUSION

The decision to break the matches for analysis of a matched pairs design and model the pairing will depend on whether the regression model is normal linear or Bernoulli logistic. With the normal linear model, the estimators of the treatment effect from the regression analysis and the matched pairs analysis are both unbiased but they have different variances. Regression turns out to be a poor alternative to matched pairs analysis (or the paired t-test) unless the number of pairs is very small and the measured covariates are able to account for a large proportion of the "information" of the omitted covariates. When the model is Bernoulli logistic, analytical results are only available asymptotically, and these indicate that logistic regression will in general be more efficient than the matched pairs analysis. For finite sample sizes, simulations suggest that logistic regression is still more favorable, but that the matched pairs analysis is preferable when the number of pairs is small, and when the regression coefficient of the omitted covariate is large and the correlation between the known covariate and the omitted covariate is small.

Table 2) Values of r_{ML}*

$\rho = 0.2$

		$\gamma = 0$		0.5		1.2	
		β_3		β_3		β_3	
m	β_2	0.3	1.2	0.3	1.2	0.3	1.2
30	0.4	1.00	1.00	0.99	0.99	0.97	0.96
	1.2	0.76	0.99	0.77	1.07	1.52	1.65
60	0.4	0.95	0.97	0.96	0.96	1.01	1.00
	1.2	0.59	0.69	0.66	0.81	1.80	1.33
100	0.4	0.95	0.96	0.96	0.99	1.35	1.09
	1.2	0.53	0.66	0.62	0.72	2.59	1.51

$\rho = 0.8$

		$\gamma = 0$		0.5		1.2	
		β_3		β_3		β_3	
m	β_2	0.3	1.2	0.3	1.2	0.3	1.2
30	0.4	1.00	1.00	1.00	1.00	0.97	1.00
	1.2	0.79	1.01	0.83	1.04	1.07	0.99
60	0.4	0.93	0.96	0.96	0.98	0.97	1.00
	1.2	0.60	0.69	0.64	0.81	0.79	1.03
100	0.4	0.95	0.97	0.96	0.99	0.96	0.98
	1.2	0.54	0.71	0.58	0.74	0.76	0.84

KEY: m ≡ number of pairs, β_2≡ regression coefficient for considering the treatment effect, β_3≡ regression coefficient for the known covariate, γ ≡ regression coefficient for the omitted covariate, and ρ ≡ correlation between x_3 and z (the omitted covariate vector).

*r_{ML} is defined as the median of 3000 replicated values of $\dfrac{(\hat{\tau}_{logit} - 2\beta_2)^2}{(\hat{\tau}_{ML} - 2\beta_2)^2}$.

Table 3) Values of r_H*

$\rho = 0.2$

		$\gamma = 0$		0.5		1.2	
		β_3		β_3		β_3	
m	β_2	0.3	1.2	0.3	1.2	0.3	1.2
30	0.4	1.00	1.00	1.00	1.00	1.00	1.00
	1.2	1.03	0.75	1.08	0.75	1.72	1.13
60	0.4	0.97	1.00	0.97	0.99	1.03	1.00
	1.2	0.59	0.83	0.68	0.89	1.71	1.45
100	0.4	0.95	0.97	0.96	0.99	1.33	1.09
	1.2	0.59	0.72	0.64	0.75	2.47	1.56

$\rho = 0.8$

		$\gamma = 0$		0.5		1.2	
		β_3		β_3		β_3	
m	β_2	0.3	1.2	0.3	1.2	0.3	1.2
30	0.4	1.00	1.02	1.00	1.04	1.00	1.00
	1.2	1.14	0.74	0.98	0.60	0.84	0.58
60	0.4	0.95	0.97	0.96	1.00	1.00	1.00
	1.2	0.62	0.76	0.67	0.81	0.89	0.98
100	0.4	0.96	0.98	0.96	0.99	0.97	0.99
	1.2	0.57	0.74	0.63	0.81	0.79	0.91

*r_H is defined as the median of 3000 replicated values of $\dfrac{(\hat{\tau}_{logit} - 2\beta_2)^2}{(\hat{\tau}_H - 2\beta_2)^2}$.

ACKNOWLEDGEMENTS

We wish to thank Professor Thomas Santner for reviewing the manuscript. This research was supported in part by Hatch grant 151-406.

REFERENCES

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research, Vol. 1: The Analysis of Case-Control Studies.* IARC Scientific Publications No. 32, International Agency for Research on Cancer, Lyon, France.

Edlefsen, L. E. and Jones, S. D. (1986). *GAUSS,* documentation version 1.00, software version 1.49B. Washington: Aptech Systems Inc.

Gail, M.; Wieand, S.; and Piantadosi, S. (1984). "Biased estimates of treatment effect in randomized experiments with non-linear regressions and omitted covariates," *Biometrika,* 71, pp. 431-444.

Gail, M.; Tan, W.; and Piantadosi, S. (1988). "Test of no treatment effect in randomized clinical trials," *Biometrika,* 75, pp. 57-64.

Greenland, S. (1986). "Partial and marginal matching in case-control studies", in *Modern Statistical Methods in Chronic Disease Epidemiology* ,(pp. 35-49) edited by Moolgavkar, S. and Prentice, R. John Wiley & Sons, Inc. , New York.

Jewell, N. P. (1984). "Small-sample bias of point estimators of the odds ratio from matched sets," *Biometrics,* 40, pp. 421-435.

Kupper, L. L.; Karon, J. M.; Kleinbaum, D. G.; Morgenstern, H.; and Lewis, D. K. (1981). "Matching in epidemiologic studies: validity and efficiency considerations," *Biometrics,* 37, pp. 271-292.

McKinlay, S. M. (1977). "Pair matching -- a reappraisal of a popular technique," *Biometrics,* 33, pp. 725-735.

Prentice, R. (1976). "Use of the logistic model in retrospective studies," *Biometrics*, 32, pp. 599-606.

Royall, R. M. (1986). "Model robust confidence intervals using maximum likelihood estimators," *International Statistical Review*, 54, pp. 221-226.

Rubin, D. B. (1973). "The use of matched sampling and regression adjustment to remove bias in observational studies," *Biometrics*, 29, pp. 185-203.

APPENDIX 1

The following tables list the simulated mean square errors of three estimators of the log odds ratio, with their respective standard errors included underneath in parentheses. $\hat{\tau}_{logit}$ is the estimator obtained from logistic regression, whereas $\hat{\tau}_{ML}$ and $\hat{\tau}_{H}$ are the two different matched pairs estimators. n1 is the number of replications in each run that led to undefined estimators for the matched pairs analysis, and n2 is the number of replications in each run that did not produce converging solutions for the logistic regression. (Note that the values for the case when m=60 have been omitted in the following tables to shorten the presentation.)

m=30

$\hat{\tau}_{ML}$	$\hat{\tau}_{H}$	$\hat{\tau}_{logit}$	n1	n2	ρ	γ	β_2	β_3
.3656	.2858	.3265	18	0	0.2	0	0.4	0.3
(.0108)	(.0076)	(.0095)						
.4857	.3625	.5039	82	14			0.4	1.2
(.0134)	(.0098)	(.0153)						
.3581	.3666	.5507	587	7			1.2	0.3
(.0085)	(.0102)	(.0167)						
.3710	.4707	.5565	880	216			1.2	1.2
(.0123)	(.0140)	(.0142)						
.4030	.3111	.3135	30	0		0.5	0.4	0.3
(.0120)	(.0086)	(.0091)						
.4694	.3496	.4755	90	15			0.4	1.2
(.0125)	(.0091)	(.0142)						
.3269	.3508	.4783	662	8			1.2	0.3
(.0086)	(.0103)	(.0151)						
.3870	.5081	.5049	893	210			1.2	1.2
(.0128)	(.0144)	(.0122)						
.4657	.3544	.2819	71	0		1.2	0.4	0.3
(.0129)	(.0094)	(.0073)						
.5151	.3762	.3852	132	2			0.4	1.2
(.0134)	(.0100)	(.0113)						

m=30

$\hat{\tau}_{ML}$	$\hat{\tau}_{H}$	$\hat{\tau}_{logit}$	n1	n2	ρ	γ	β_2	β_3
.3544	.4484	.5308	812	2	0.2	1.2	1.2	0.3
(.0111)	(.0129)	(.0117)						
.4122	.5864	.5601	1028	84			1.2	1.2
(.0141)	(.0156)	(.0123)						
.3804	.2979	.3296	20	0	0.8	0	0.4	0.3
(.0112)	(.0079)	(.0090)						
.4698	.3491	.4877	74	13			0.4	1.2
(.0126)	(.0091)	(.0139)						
.3353	.3341	.5397	596	6			1.2	0.3
(.0081)	(.0098)	(.0169)						
.3757	.4739	.5444	798	195			1.2	1.2
(.0119)	(.0136)	(.0136)						
.4071	.3185	.3544	20	0		0.5	0.4	0.3
(.0118)	(.0087)	(.0103)						
.5371	.3931	.5970	174	62			0.4	1.2
(.0139)	(.0103)	(.0160)						
.3292	.3667	.5291	647	38			1.2	0.3
(.0087)	(.0105)	(.0159)						
.4567	.6418	.5054	1022	459			1.2	1.2
(.0155)	(.0169)	(.0129)						
.4957	.3707	.4318	83	7		1.2	0.4	0.3
(.0127)	(.0092)	(.0120)						
.6049	.4512	.6518	363	223			0.4	1.2
(.0169)	(.0132)	(.0176)						
.3883	.5120	.5144	891	127			1.2	0.3
(.0127)	(.0143)	(.0121)						
.6469	.9583	.5662	1374	808			1.2	1.2
(.0228)	(.0235)	(.0174)						

m=100

$\hat{\tau}_{ML}$	$\hat{\tau}_{H}$	$\hat{\tau}_{logit}$	n1	n2	ρ	γ	β_2	β_3
.1011	.0950	.0903	0	0	0.2	0	0.4	0.3
(.0029)	(.0026)	(.0023)						
.1469	.1341	.1304	0	0			0.4	1.2
(.0045)	(.0039)	(.0035)						
.2676	.2048	.1259	13	0			1.2	0.3
(.0091)	(.0058)	(.0036)						
.3187	.2287	.1953	37	0			1.2	1.2
(.0095)	(.0057)	(.0064)						

m=100

$\hat{\tau}_{ML}$	$\hat{\tau}_{H}$	$\hat{\tau}_{logit}$	n1	n2	ρ	γ	β_2	β_3
.1123	.1047	.0877	0	0	0.2	0.5	0.4	0.3
(.0035)	(.0031)	(.0023)						
.1443	.1316	.1211	0	0			0.4	1.2
(.0042)	(.0037)	(.0032)						
.2700	.2054	.1232	17	0			1.2	0.3
(.0089)	(.0056)	(.0031)						
.3268	.2369	.1972	61	0			1.2	1.2
(.0096)	(.0057)	(.0061)						
.1368	.1256	.1014	0	0		1.2	0.4	0.3
(.0044)	(.0038)	(.0024)						
.1679	.1508	.1156	0	0			0.4	1.2
(.0053)	(.0044)	(.0029)						
.3281	.2340	.3439	45	0			1.2	0.3
(.0098)	(.0058)	(.0056)						
.3527	.2471	.2837	89	0			1.2	1.2
(.0094)	(.0057)	(.0059)						
.0979	.0925	.0882	0	0	0.8	0	0.4	0.3
(.0029)	(.0026)	(.0023)						
.1478	.1345	.1315	0	0			0.4	1.2
(.0050)	(.0042)	(.0037)						
.2604	.1982	.1220	13	0			1.2	0.3
(.0087)	(.0055)	(.0036)						
.3231	.2298	.2032	38	0			1.2	1.2
(.0097)	(.0057))	(.0062)						
.1089	.1011	.0925	0	0		0.5	0.4	0.3
(.0039)	(.0033)	(.0025)						
.1834	.1631	.1641	0	0			0.4	1.2
(.0056)	(.0046)	(.0044)						
.2858	.2115	.1418	22	0			1.2	0.3
(.0093)	(.0058)	(.0044)						
.3565	.2534	.2753	101	14			1.2	1.2
(.0094)	(.0056)	(.0090)						
.1533	.1389	.1154	0	0		1.2	0.4	0.3
(.0048)	(.0041)	(.0031)						
.2789	.2322	.2266	0	0			0.4	1.2
(.0089)	(.0067)	(.0071)						
.3304	.2333	.1805	54	1			1.2	0.3
(.0097)	(.0056)	(.0049)						
.3624	.2697	.3552	223	45			1.2	1.2
(.0078)	(.0059)	(.0096)						

APPENDIX 2: SIMULATION DETAILS

The GAUSS program was run on a IBM personal system/2™ Model 60 machine. 3000 replications were performed for each combination of the five parameter values: m, ρ, γ, β_2, and β_3. When calculating $\hat{\tau}_{logit}$, replications that did not give converging solutions within 30 iterations of a Newton-Raphson algorithm were discarded. (To ensure that 30 is a reasonable choice, a subset of the data sets that did not converge within 30 iterations were inspected and these were all verified to yield nonconvergent solutions.) For the matched pairs analysis, a replication was discarded when $\hat{\tau}_{ML}$ was undefined.